Capstone Project#1

Xiao Xi

Introduction

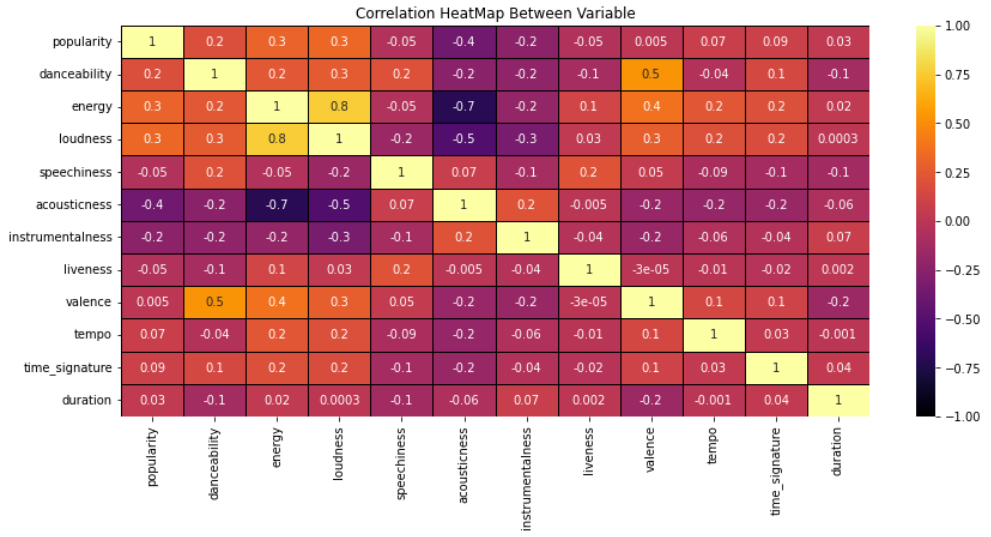
In this project, I will use Python to analyze the datasets from Spotify research. I downloaded two csv files: tracks.csv and SpotifyFeatures.csv from the website as the datasets for this project. From these datasets, I am interested in finding the correlations between the variables, and find the most popular songs in different dimensions. The hypothesis of these datasets is the louder the music is, the more energy rating the song has. Another hypothesis is that popularity is negatively related to acousticness. Below is the process of how I substantiate the hypothesis with Python.

Import the datasets and conduct preliminary analysis

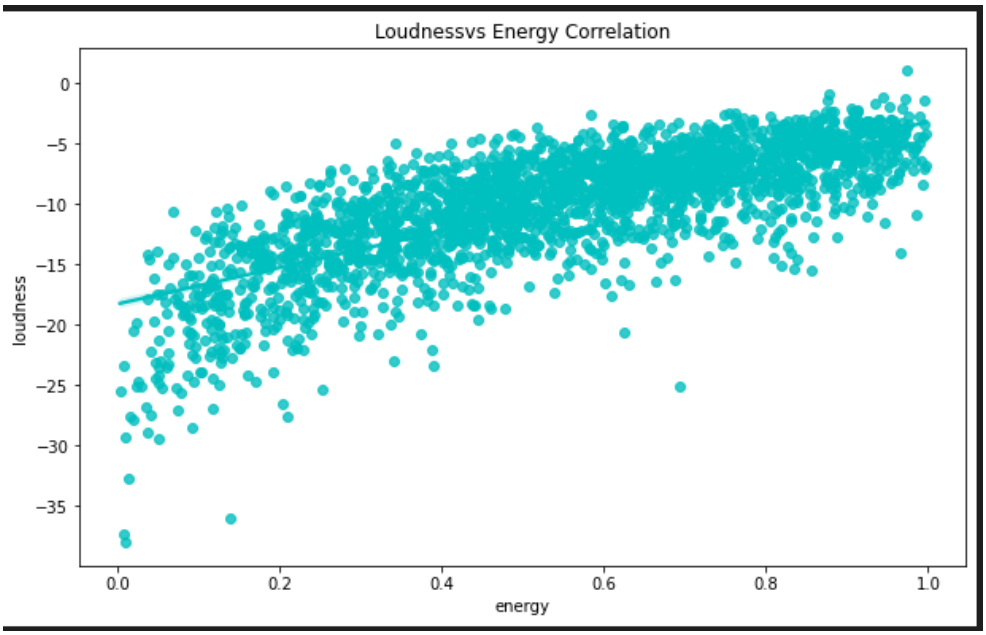
First of all, I imported several Pandas modules to get ready for the upcoming analysis. Then I imported the tracks.csv file from my desktop. After I set up the dataframe, I used sort functions to get the most popular songs based on certain criteria and the basic statistical measures of the datasets (i.e. mean, sum, deviation,etc.). With this preliminary analysis, I can start visualize the data using graphs and diagrams.

Data visualization

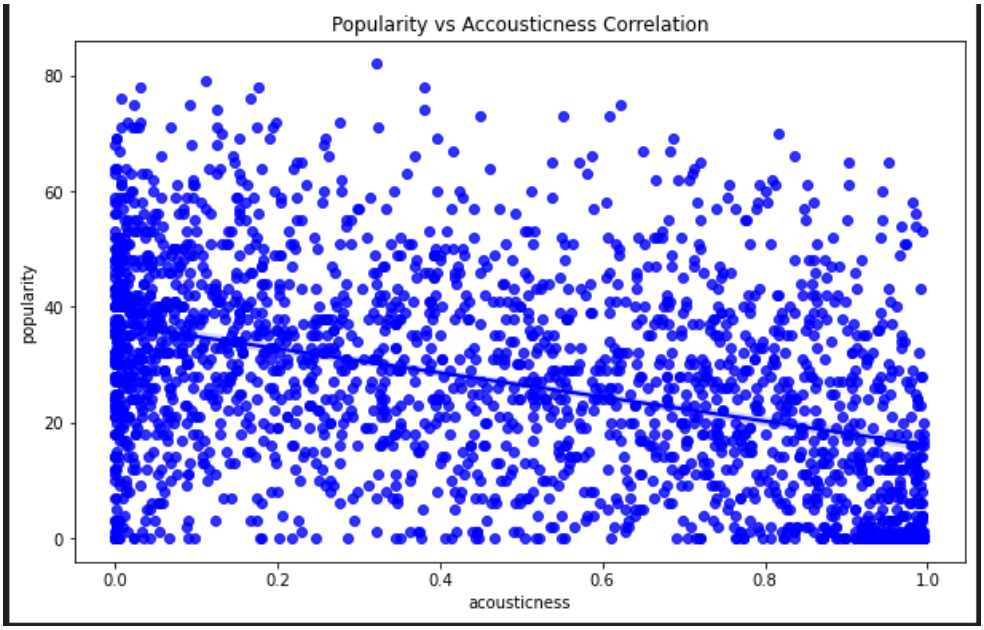
To visualize the datasets, I used the heatmap first to analyze the correlation coefficient between each variable per below:



From the above heatmap, we can see that the darker the filling, the stronger relationship between the variable is. Next, I made a plot to iterate the relationship between loudness and energy as follows:

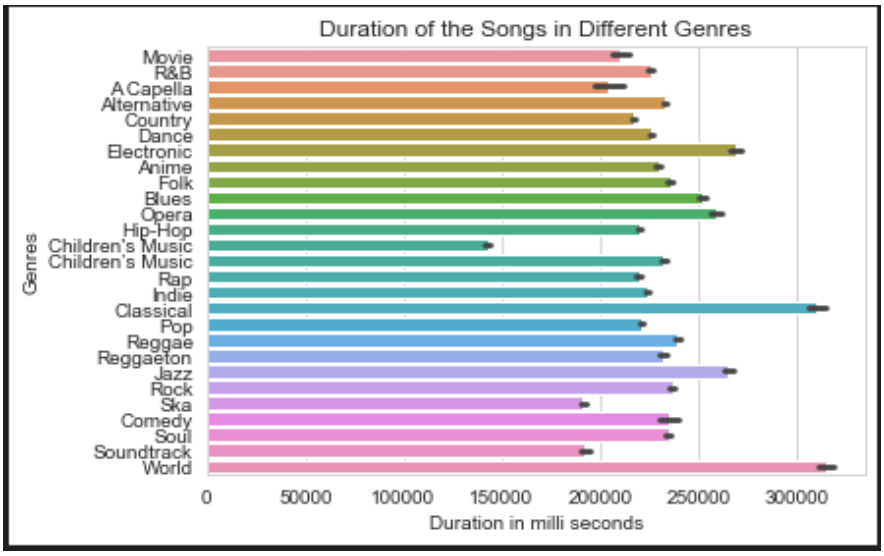


From the graph above, it looks like the loudness is positively related to energy, which proves my hypothesis. Then I ran a similar analysis between popularity and acousticsness per below:

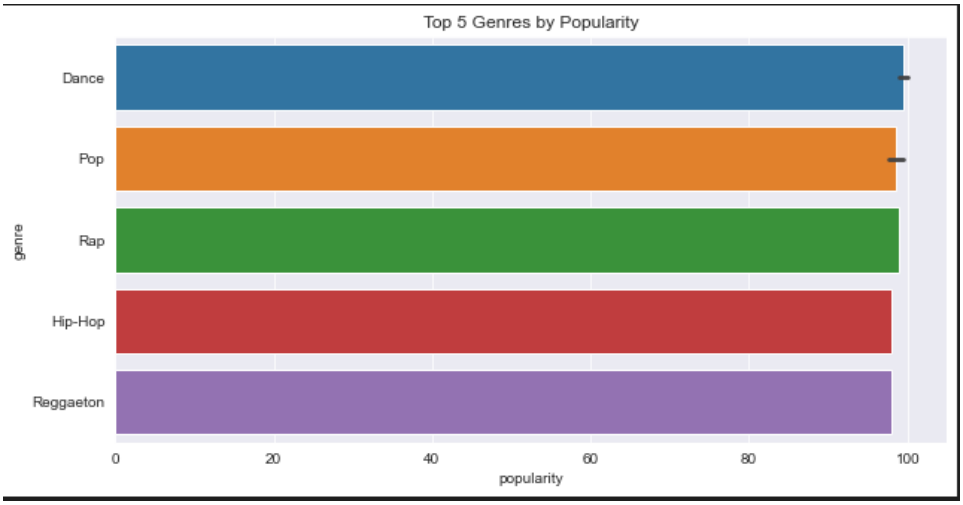


From this graph, it seems that the relationship is weak but there's a slight trend that the more acousticness is the less popularity the sone will be.

Lastly, I imported the SpotifyFeatures.csv file, and drew a graph to show the duration of songs in different genres per below:



Also, I plotted another diagram showing the top 5 genres by popularity:



Conclusion

In summary, it looks like my hypothesis is correct, and there could be more combinations of relationships to be discovered by using Pandas in Python.